Title: Origin of horsetails and the role of whole genome duplication in plant macroevolution

James W. Clark[1,2*], Mark N. Puttick[2,3] & Philip C.J. Donoghue[2]

[1]Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB Oxford, United Kingdom.
[2]School of Earth Sciences University of Bristol, BS8 1TQ Bristol, United Kingdom
[3]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, BA2 7AY Bath, United Kingdom
*****Author for correspondence

Summary

Whole Genome Duplication (WGD) has occurred commonly in land plant evolution and it is often invoked as a causal agent in diversification, phenotypic and developmental innovation, as well as conferring extinction resistance. The ancient and iconic lineage of *Equisetum* is no exception, where WGD has been inferred to have occurred prior to the Cretaceous-Paleogene (K-Pg) boundary, coincident with WGD events in angiosperms. In the absence of high species diversity, WGD in *Equisetum* is interpreted to have facilitated the long-term survival of the lineage. However, this characterisation remains uncertain as these analyses of the *Equisetum* WGD event have not accounted for fossil diversity. Here we analyse additional available transcriptomes and summarise the fossil record. Our results confirm support for at least one WGD event shared among the majority of extant *Equisetum* species. Furthermore, we use improved dating methods to constrain the age of gene duplication in geological time and identify two successive *Equisetum* WGD events. The two WGD events occurred during the Carboniferous and Triassic, respectively, rather than in association with the K-Pg boundary. WGD events are believed to drive high rates of trait evolution and innovations, but analysed trends of morphological evolution across the historical diversity of *Equisetum* provide little evidence for further macroevolutionary consequences following WGD. WGD events cannot have conferred extinction resistance to the *Equisetum* lineage through the K-Pg boundary since the ploidy events occurred hundreds of millions of years before this mass extinction and we find evidence of extinction among fossil polyploid *Equisetum* lineages. Our findings precipitate the need for a review of the proposed roles of WGDs in biological innovation and extinction survival in angiosperm and non-angiosperm lineages alike.

1. Introduction

The prevalence of Whole Genome Duplication (WGD) in land plants has contributed to the widely held view that WGD is an agent of macroevolutionary change [1]. The most striking pattern to have emerged is the apparent temporal clustering of WGD events about the Cretaceous-Palaeogene (K-Pg) boundary interval [2-4]. Perhaps inevitably, this has led to suggestions that WGD facilitated the survival and success of plant lineages in the wake of the attendant ecological disturbance and mass extinction [5-7]. Further, polyploid formation at mass extinction events is predicted to have been higher, as environmental disturbance and stress led to the formation of unreduced gametes [8, 9]. However, the WGD-K-Pg hypothesis is dependent on the accuracy and precision of estimates for the timing of WGD events.

Transcriptomics of *Equisetum giganteum* have revealed that, like many other land plant lineages, *Equisetum* underwent at least one round of WGD [10]. The phylogenetic position of *Equisetum* on a long depauperate branch makes direct molecular dating challenging and hence previous studies have broad confidence intervals around estimated ages. Nevertheless, age estimates from synonymous substitutions (*Ks*) between duplicate gene pairs have been interpreted cautiously to reflect a duplication age overlapping the K-Pg boundary [10].

WGD is often proposed as a driver of species diversification [11]. *Equisetum* seems to be an exception, as with only 15 extant species the genus hardly evidences a link between WGD and diversification. In lieu of high species diversity, Vanneste *et al.* [10] have suggested that the WGD event may have contributed to the longevity of the lineage, despite estimating a relatively recent *Equisetum* WGD. WGD is also generally proposed as a driver of phenotypic innovation [12], however, few studies consider the diversity of extinct forms in the context of WGD [13]. This is pertinent to *Equisetum* which exhibits a rich evolutionary history that has been revealed by several recent palaeontological discoveries [14-17].

To test the association of *Equisetum* WGD and the K-Pg extinction event, we present a thorough analysis of the timing of WGD within Equisetales and its putative macroevolutionary consequences. We refine the phylogenetic position of putative WGD events and use molecular clock methods to show that WGD occurred well before the K-Pg, closer in age to the more ancient and profound Permian-Triassic extinction event. Further, we show that the WGD is not responsible for the phenotypic distinctiveness of *Equisetum.* There is no evidence that WGD conferred extinction resistance to Equisetales with many Mesozoic lineages not making it through the K-Pg mass extinction.

2. Materials and Methods

(a) Transcriptome Assembly

Assembled transcriptomes were collected from the 1KP dataset for *Equisetum diffusum, Equisetum hyemale, Culcita macrocarpa, Ophioglossum petiolatum, Tmesipteris parva, Selaginella kraussiana, Danaea nodosa* and *Botrypus virginianus*, and an additional transcriptome for *Equisetum giganteum* was obtained from [10].

Paired end short reads were downloaded from the SRA archive for *Equisetum arvsense* (SRR4061754), *Equisetum telmateia* (SRR4061752) and *Equisetum ramossisimum* (SRR5499399), and assembled following [18]. Reads were trimmed of adapter sequences using Trimmomatic v.0.35 [19] using default settings. Assembly was performed using Trinity [20] using default settings. Redundant transcripts were removed using CD-HIT with a cluster value of 0.95 [21]. Each transcript was converted into the single best amino acid sequence using TransDecoder [22]. The assembly of the *E. arvense, E. ramosissimum* and *E. telmateia* transcriptomes after clustering resulted in 24,187, 58,549 and 61,969 transcripts.

(b) *Ks* analysis

We compared rates of synonymous substitution between paralogous genes in *E. hyemale* and *E. diffusum,* that represent the subgenera *Hippochaete* and *Equisetum,* respectively. Analyses were performed using default parameters and the 'phyml' node-weighting method in the *wgd* package [23-26]. *Ks* distributions were plotted based on node-averaged values as calculated in the *wgd* package. Gaussian mixture models were fitted to the *Ks* distribution following the *wgd* pipeline, with the optimal number of components assessed using the Bayesian Information Criterion (BIC).

(c) Gene family assignment

Orthogroups from the transcriptomes were inferred using Orthofinder v.2.2.6 [27] under a Diamond sequence search. The Orthofinder analysis initially produced 27,038 orthogroups. An initial filtering step was performed to remove orthogroups that did not contain at least one representative from 75% of species. Remaining orthogroups were aligned using MUSCLE and trimmed using trimal [28]. A second filtering step removed all alignments shorter than 200 amino acids, resulting in 5,009 orthogroups. Phylogenetic inference was performed on each remaining orthogroup under the best-fitting model and maximum likelihood criterion in IQ-TREE [29], with 1000 ultra-fast bootstrap replicates [30].

(d) Species Divergence Time Estimation

Single copy orthogroups from the Orthofinder output formed the basis of a dating analysis. An alignment of 45,977 amino acids was partitioned by gene for a topology search using the edge-linked option (-spp) in IQ-TREE [29].

The topology formed the basis of a fixed-topology node-calibrated molecular clock analysis in MCMCtree [24]. Node calibrations were specified with a uniform distribution spanning the hard minimum and soft maximum constraints (with a 2.5% tail distribution) established using *MCMCtreeR* in R (Table 1) [31]. Previous studies have placed the fossil taxon *Equisetum fluviatoides* as sister to *E. diffusum* [17]. However, our analyses supported a *E. fluviatoides* as sister to both *E. diffusum* and *E. arvense*, and so we established a calibration for the divergence of the two subgenera (Supplementary Methods). The mean rate was assigned a gamma prior, determined based on the mean number of substitutions along the tree scaled by the approximate geological age, with a total of 0.12 substitutions per site per million years. To ensure the model sampled from this distribution we fixed the shape parameter to two and adjusted the scale parameter to 16 [32, 33]. The analysis was run without sequence data to ensure that the effective time priors were compatible with the palaeontological and phylogenetic constraints informing the specified node calibrations [34]. Using the approximate likelihood method [35], we ran two independent analyses, each for 5,000,000 generations, discarding the first 1,000,000 generations as burn-in. Convergence of each run was assessed using Tracer [36].

(e) Gene tree and species tree reconciliation

Gene trees inferred from Orthofinder were reconciled with the dated species tree. Gene trees were inferred under a DTL (Duplication, Transfer, Loss) model using a maximum likelihood criterion in ALE (Amalgamated Likelihood Estimation) [37]. The reconciliations were performed using 1000 ultrafast bootstrap replicates as tree samples. As there is no prior hypothesis regarding an ancient hybridization (allopolyploidy) event in *Equisetum*, we set a low prior rate of gene transfer (0.1). The total number of duplications was summed for each branch in the phylogeny based on the number of inferred duplications across each of the 1000 sampled trees for each gene family.

(f) Dating whole genome duplication

Gene families inferred to have duplicated along the branch leading to *Equisetum* were sampled from the ALE output (Supplementary Fig S1). To evaluate the hypothesis of a single WGD event in *Equisetum,* we selected gene families that contained a single duplication along this branch for a molecular clock analysis. Following [38], gene families were used if they: (i) had a clear topological signal of the WGD event, represented by two paralogous copies present in all *Equisetum* species forming two monophyletic groupings; (ii) had a topology congruent with current understanding of tracheophyte phylogeny; and (iii) did not have a signal of additional duplication events within *Equisetum.* We conducted a molecular clock analysis for each gene family with the same settings as used for the species divergence estimation. The 95% Highest Posterior Densities (HPDs) were combined between all gene families. Peaks in this combined posterior distribution may represent duplication events common to multiple gene families. To determine which gene families coincide with each peak, the peaks in the combined posterior distribution were described using Gaussian mixture models (GMMs) and the overlap between these peaks and the individual gene posterior distributions were estimated using an overlapping coefficient [39]. Gene families with an overlap > 0.8 for each respective peak were selected and concatenated. Molecular clock analyses were performed for families corresponding to each peak, with the same set of fossil calibrations employed as in the species divergence time estimation, with the exception that the calibration within *Equisetum* was cross-calibrated on both sides of the duplication. Analyses were performed as for the species divergence estimation.

To consider the possibility of multiple WGD events, we repeated the analysis with gene families containing at least two duplications (four copies of each gene) in all extant *Equisetum* species, allowing for simultaneous age estimation of two duplication nodes.

(g) Dating of Fossils and Extant Taxa

We used previously assembled phenotypic and molecular matrices of 77 binary and multistate phenotypic characters and the *rbcL, atpA, atpB* and *matK* chloroplast genes [17]. The matrix contained 49 taxa, including 17 extant and 32 fossil taxa spanning the Sphenophyllales + Equisetales as well as outgroup taxa *Hamatophyton verticillatum*, *Rotafolia songziensis*, *Ophioglossum reticulatum* (Ophioglossales) and *Psilotum nudum* (Psilotales).

We estimated divergence times using the estimates obtained from the molecular species divergence analysis as priors on nodes present in this dataset. Fossil tip ages were based on a uniform distribution across their occurrence ranges (Supplementary Table 1) and a

uniform distribution was placed on the root between 451-384 million years [33]. A stepping stone analysis was used to test for the best-fitting clock model in MrBayes v.3.2.6 [40, 41]; this showed significant support for the correlated model [42] over the Independent Gamma Rates [43] and strict clock models. A correlated rates clock model [42] was implemented with the clock rate prior set as a lognormal distribution; the mean of the lognormal distribution was estimated from a topological analysis to estimate the tree height scaled by the approximate geological age of the root (0.02 substitutions $site^{-1}$ million $years^{-1}$) [44]. Finally, we set a uniform birth-death prior across the tree [41]. The phenotypic data and each gene were partitioned separately, with molecular data analysed under the GTR+Γ model and the phenotypic data under the MKv+ Γ model [45]. Four independent chains were run for 20,000,000 generations. Convergence between the chains was assessed based on the average standard deviation of split frequencies (< 0.01), Effective Sample Size (target > 200) and by examining the parameters of the chain in Tracer [36].

(h) Rates of Phenotypic Evolution

To examine the rates of phenotypic evolution across the tree, we performed a morphological clock analysis using only the phenotypic dataset with the tree constrained to the topology resolved by the combined analysis. A relaxed clock model was used, allowing rates to vary between branches.

The rate of phenotypic evolution was estimated by sampling the effective branch lengths from 1000 points of the posterior distribution; the mean rates were estimated from these samples. Only branches from the majority-rule consensus topology were considered for further analyses; from the 1000 posterior samples, rates were summarised for branches on the posterior tree that matched branches on the majority-rule consensus tree.

(i) Phenotypic Disparity

The phenotype matrix was recoded following [46], such that non-applicable (NA) states were coded as ‘0’ and missing data as ‘?’, to distinguish the two types of ‘missing data’ [47]. The distance between taxa was calculated using Gower’s dissimilarity metric [48]. The distances were projected into two-dimensional space using Non-metric Multi-Dimensional Scaling (NMDS). We plotted a phylomorphospace using the majority-rule (50%) consensus tree from the total evidence analysis [49]. The most likely ancestral state was reconstructed along the tree by summarising states across 1000 stochastic character maps [50]; the estimated states were used to position the nodes within the morphospace.

We calculated mean disparity as Sum Of Variances from the distance matrix [51] using *dispRity* in R [52]. Disparity through time was estimated using the time-slicing approach using 10 bins and the 'gradual split' model as implemented in *dispRity*, with the probability of a character state being that of either the descendent or the ancestor dependent on the length of the branch [52].

(j) Genome Size Analysis

Genome size estimates (1C-values) were downloaded from the c-value database [53]. The 1C-values were estimated for fossil taxa by Franks *et al.* [54] who derived a linear regression model for the relationship between 1C-value and stomata guard cell length. They estimated 1C-value for members of Sphenophyllales (*Sphenophyllum)* and Calamitaceae (*Calamocladus*) as well as *Equisetum haukeanum*. For this analysis we took the values for Sphenophyllales and Calamitaceae to be representative of each lineage. We used the linear model ($y = 1.83x + -5.46$) to convert the logged guard cell widths of other fossil *Equisetum* and to a logged 1C-value [14-16, 54, 55]. In total, 21 1C-values were obtained (Supplementary Table 1) and were analysed as continuous characters in BayesTraits v.3 [56] using a homogeneous continuous random walk model and the ancestral 1C-values were estimated at internal nodes. The MCMC was run for 15,000,000 generations, with the first 10,000,000 generations discarded as burn-in.

3. Results

(a) Transcriptomic Analyses Reveal Triassic and Carboniferous WGD Events

The distribution of *Ks* values in *E. hyemale* and *E. diffusum* exhibit at least 3 conspicuous peaks: one close to 0.1 representing recent duplicates, another with a mean close to 1, and third more ancient peak close to 2 (Fig 1). Mixture modelling supported 4 components, but the fourth component had a low mean weight (Fig 1, Supplementary Fig S1). Coincidence of these peaks suggests that the WGD event initially identified in *E. giganteum* is shared between both subgenera, though *Ks* values >2 are increasingly unreliable predictors of WGD [57].

ALE analysis revealed rates of duplication that were generally higher on terminal branches (likely due to recent local duplication events) and some of the long branches included in the study. Among all branches, however, ALE provided strong support for a duplication event on the branch leading to *Equisetum* (Supplementary Fig S2). 240 gene families were selected from the ALE output that showed a clear signal of the duplication

event. Molecular clock analyses of these gene families supported two clear clusters of ages (Fig 2). For each cluster, we found 52 and 51 corresponding gene families that were concatenated to form alignments of 21,894 and 19,360 amino acids. These analyses suggested a first duplication within the interval 329-307 Ma (Serpukhovian-Moscovian: mid-late Carboniferous) and a second within 253-233 Ma (Changhsingian-Carnian: latest Permian to Late Triassic) (Fig 3).

We identified a further 14 gene families with a clear signal of two successive duplications with all 4 paralogs retained. The two successive duplications were estimated to 360-322 Ma (Fammenian-Bashkirian: latest Devonian to mid Carboniferous) and 261-211 Ma (Capitanian-Norian: late Permian to Late Triassic; Supplementary Fig S3).

(b) An Evolutionary Framework: Triassic-Jurassic origin of total-group *Equisetum*

Analysis of the combined molecular and morphological dataset partially resolved the backbone phylogeny of Equisetales (Fig 4). Monophyly of Equisetales is strongly supported, with Neocalamitaceae as sister to all remaining Equisetaceae, but there is only weak support for Neocalamitaceae. As with [17], we resolve *Equisetites arenaceus* and *Spaciinodum collinsonii* as sister to the total group *Equisetum*.

Relationships within *Equisetum* are poorly resolved; the two subgenera (*Equisetum* and *Hippochaete*) are well supported, as are the positions of *E. clarnoi* and *E. fluviatoides* within each, respectively. The relationships of the outgroups are also poorly resolved, including the order of divergence of Archaeocalamitaceae and Calamitaceae, although as we confirm that Equisetaceae did not originate from within Calamitaceae.

We estimate a Devonian origin of both sphenopsids and ferns. Sphenophyllales and Equisetales diverged during the Carboniferous along with most of the extinct lineages of Equisetales, including the Archaeocalamitaceae and Calamitaceae. Equisetaceae and Neocalamitaceae diverged during the Permian. We report a Triassic-Jurassic origin of total group *Equisetum*, but a Cretaceous origin of the crown-group, with both extant subgenera originating during the Palaeogene (Supplementary Fig S4).

(c) High Rates of Phenotypic Evolution at The Origin of Major Clades

Rates of phenotypic evolution are heterogeneous across the tree (Fig 4). The origin of major lineages is marked by the fastest rates of phenotypic evolution, including Equisetales, Equisetaceae and *Hippochaete* (Fig 4). Generally, phenotypic evolution is much greater between higher-order lineages than within them, with slow rates observed within

Equiseteceae and most lineages within Calamitaceae, except the branch leading to *Cruciaetheca*.

High rates of phenotypic evolution correspond to large distances in morphospace (Fig 5a). Major lineages cluster tightly within morphospace across both axes, though on the individual axes there is considerable overlap. The proportion of total disparity represented by extant taxa is low (Fig 5b) and disparity through time analyses show that modern levels of disparity are a small fraction of a Carboniferous acme (Fig 5c). Mean disparity, measured as the average Euclidean pairwise distance between taxa, is lower in Equisetaceae (0.195) than Calamitaceae (0.381), but they do occupy a novel region of morphospace.

(d) Genome Duplication and Genome Size

Reconstruction of ancestral genome size within Sphenopsida reveals that the largest genome sizes are found within extant *Equisetum* (mean ancestral 1C-value = 17.09pg), in particular the subgenus *Hippochaete* (ancestral 1C-value = 20.9pg) (Fig 6). Across nodes, we observed three large increases in genome size: from the base of *Equisetum* to *Hippochaete* (17.6pg to 20.9pg), from the base of Equisetales to total group *Equisetum* (3.9pg to 11.01pg), and from total group to crown group *Equisetum* (11.01 to 17.6pg) (Fig 6).

4. Discussion

(a) Duplication and Evolution in *Equisetum*

The WGD shared by extant *Equisetum* was previously proposed as one of several WGD events that coincide with the K-Pg boundary [2, 10]. The significance of this clustering of events has been explored from various angles: that WGD confers an ‘extinction resistance’, that WGD may have provided a means of rapid adaptation amidst ecological disturbance, that WGD may be a response to environmental stresses, and that WGD itself might just be a non-selective consequence [58] of a switch to vegetative reproduction often associated with polyploidy [2, 59, 60]. The new age estimates presented here render these hypotheses unlikely given that the WGDs predate the K-Pg mass extinction by hundreds of millions of years. Indeed, we find no evidence of beneficial evolutionary consequences of WGD in *Equisetum*, suggesting that these events do not universally precipitate changes on the macroevolutionary scale across the tree of life.

Our analyses supported multiple bursts of gene duplication throughout the evolution of the *Equisetum* lineage. Their interpretation as WGD events can be difficult [61], yet their clustering within time and the repeated history of WGD across land plants suggests that there

is a high probability that they represent WGD events. Though congruent with the findings of Vanneste *et al.* [10], we have better resolved the phylogenetic position of these putative WGD events and find that they are likely shared by both subgenera of *Equisetum* (Fig 1). However, the WGD event proposed by Vanneste *et al.* [10] to have occurred in *E. giganteum* was known only from a single transcriptome and the geological age was difficult to constrain using both phylogenomic and *Ks* methods. Indeed, ages inferred directly from *Ks* distributions can be inaccurate due to sequence saturation and the assumption of a strict clock [57, 62].

Using phylogenomic and molecular clock methods, we estimated both events to have occurred long before the K-Pg boundary. Rather, these WGD events are among the most ancient detected in land plants, occurring within the latest Devonian-mid Carboniferous and late Permian-Late Triassic, respectively (Fig 3). This estimate is comparable in precision to recent estimates for other WGD events associated with the K-Pg boundary [63] and serves to highlight the power of these methods to constrain the timing of the event to within 20 million years, along one of the most isolated branches within living land plants. The discrepancy in age for the *Equisetum* WGD events reported here and by Vanneste *et al.* [10] may be due to the initial paucity of transcriptomic data representative of the lineage and highlights the benefits of increased taxonomic sampling and the value of concatenation in estimating the timing of WGD events [1].

We reconstructed the evolutionary history of Equisetales using a combination of molecular and phenotype data in a Bayesian framework (Fig 4). Broadly, the relationships resolved are congruent with previous parsimony-based results [17], though the species relationships are less well resolved. The lack of resolution in the phylogeny here may be the consequence of the previously-used parsimony methods producing more highly-resolved, but less accurate trees compared to Bayesian analyses of morphological data [64, 65]. Nevertheless, our results corroborate the distinction between the Calamitaceae and Equisetaceae and the hypothesis that both lineages have evolved independently since the Carboniferous (Fig 4).

Crucially, these analyses provide a framework in which WGD can be considered in light of both extant and extinct diversity. We have shown that the more ancient WGD event took place prior to the divergence of Equisetaceae and Neocalamitaceae, and the more recent WGD event appears to coincide with the origin of Equisetaceae, either prior to, or after the divergence of *Spaciinodum*. As well as a establishing a more precise estimate for the timing

of WGD, our analyses place WGD within the context of the gross historical diversity of the lineage, rather than merely the net diversity that has survived to the present. This represents a novel approach to understanding the role of WGD in land plant evolution that is likely to be key to more thoroughly testing existing hypotheses, such as the proposed link between WGD events and the K-Pg mass extinction event in angiosperm evolution.

(b) Evolutionary consequences of WGD in a non-angiosperm lineage

The ancient timing of the *Equisetum* WGD events could be interpreted to strengthen the hypothesis that WGD has facilitated the longevity of the lineage [10]. The tentative hypothesis that the *Equisetum* WGD event conferred extinction resistance across the K-Pg seems unlikely given our estimates for the timing of the WGD events, and current hypotheses linking WGD to success emphasize only short-term advantages. Furthermore, our analyses have shown that many polyploid taxa descended from the WGD events are now extinct.

WGD events have also been implicated as drivers of phenotypic variance within the plant kingdom. Multiple models and a few examples demonstrate how novel traits have arisen in the wake of WGD that have been maintained and diversified on a macroevolutionary scale [12, 66]. The precise estimates that we have obtained for the timing of the WGD events allow us to constrain them within tight bounds on the species phylogeny and to consider their impact within the context of subsequent phenotypic evolution. The evolution of Equisetales is generally associated with relative stability and few character state changes, yet the first WGD event coincides with higher rates of phenotypic evolution (Fig 4) and each WGD event also coincides topologically with a movement into a novel area of morphospace (Fig 5a).

However, extant *Equisetum* and the fossil taxa that descended from the WGD event represent only a fraction of the phenotypic diversity of Equisetales (Fig 5b). In addition, both Equisetales and Calamitaceae exhibit fast early rates of phenotypic evolution (Fig 4); Calamitaceae also achieved greater disparity (Fig 3a). Indeed, while WGD may have played a role in promoting phenotypic novelty, it has not been sufficient to sustain disparity over time (Fig 3c). Based on previously identified synapomorphies [17], the first WGD event coincides with the evolution of lacunae (vallecular canals), the loss of internode differentiation, alternating sporangiophore shields, an increase in sporangium numbers and, possibly, the expression of all three reproductive regulatory modules [17]. The second WGD also coincides with a number of synapomorphies, including alternating ribs, leaf tips, and a reduction in the length of reproductive structures [17]. Throughout the evolutionary history of

Equisetales, the accumulation and transformation of characters associated with the extant taxa is gradual and many of the distinguishing features, including a compacted strobilus and small size, have evolved slowly and in a mosaic pattern over several nodes [17, 67, 68]. This suggests that while WGD may have had a role in promoting the diversity of the Equisetaceae, it was not a prerequisite to the evolution of disparity within Equisetales.

(c) Genome size correlates with WGD in *Equisetum*

Genome size evolution within Equisetales shows that the inferred WGD events may also correlate with an increase in ancestral genome size (Fig 6). This is in some ways surprising since the signal of genome duplication in genome size estimates rapidly erodes across most plant genomes [69, 70]. However, there is also a more recent shift towards much larger genomes that does not appear to be associated with a WGD event (Fig 6). As there are no extant members of Calamitaceae it is not possible to rule out the possibility that they may have undergone their own independent WGD event. However, the small genome size inferred for Calamitaceae [54] and relative stasis of fern genome evolution means that we may speculate that there may have been no further WGD events in this lineage [71]. Multiple WGD events may in part explain the fixed high chromosome numbers shared among extant species of *Equisetum* [71], yet does not appear to explain the distribution of genome sizes between the two extant subgenera.

Clearly, to elucidate a macroevolutionary role for WGD in land plant evolution, it is insufficient to consider only extant taxa. *Equisetum* is a good example, since its extant diversity is a poor representation of the taxonomic and phenotypic diversity that existed historically within Sphenopsida. Here, we suggest that a combination of palaeontological and genomic approaches provides additional power and greater insight when considering the impact of ancient or ‘palaeo’-polyploidy.

5. Conclusions

It is generally accepted that WGD events are agents of macroevolutionary change. Here, we have shown that a combination of macroevolutionary and comparative genomic approaches can be used to improve estimates of the timing and characterise outcomes of WGD. In *Equisetum,* WGD did not coincide with the K-Pg boundary, nor does it appear to have facilitated greater resistance to extinction. Rather, while WGD in *Equisetum* appears to correlate with the occupation of novel regions of morphospace, it has not led to significant morphological diversification. The formative role of WGD in the evolutionary history of

many angiosperm lineages is generally accepted, yet its role remains to be explored in many other plant lineages where rates of WGD are expected to be high. It is possible that differing genome dynamics may determine equally different roles for WGD in macroevolution.

Acknowledgements

The authors thank members of the Bristol Palaeobiology Group, Jill Harrison, and Andrew Leitch for helpful discussion. JC was funded by a BBSRC SWBIO DTP studentship, MNP by an 1851 Research Fellowship from the Royal Commission for the Exhibition of 1851, PCJD by NERC (NE/N002067/1) and BBSRC (BB/N000919/1).

References

1. Clark J.W., Donoghue P.C.J. 2018 Whole-Genome Duplication and Plant Macroevolution. *Trends in Plant Science*. (doi:10.1016/j.tplants.2018.07.006).
2. Lohaus R., Van de Peer Y. 2016 Of dups and dinos: evolution at the K/Pg boundary. *Current Opinion in Plant Biology* **30**, 62-69. (doi:10.1016/j.pbi.2016.01.006).
3. Vanneste K., Maere S., Van de Peer Y. 2014 Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**(1648). (doi:10.1098/rstb.2013.0353).
4. Koenen E.J., Ojeda D.I., Steeves R., Migliore J., Bakker F.T., Wieringa J.J., Kidner C., Hardy O., Pennington R.T., Herendeen P.S. 2019 The Origin and Early Evolution of the Legumes are a Complex Paleopolyploid Phylogenomic Tangle closely associated with the Cretaceous-Paleogene (K-Pg) Boundary. *bioRxiv*, 577957.
5. Renne P.R., Sprain C.J., Richards M.A., Self S., Vanderkluysen L., Pande K. 2015 State shift in Deccan volcanism at the Cretaceous-Paleogene boundary, possibly induced by impact. *Science* **350**(6256), 76-78.
6. Wilf P., Johnson K.R. 2004 Land plant extinction at the end of the Cretaceous: a quantitative analysis of the North Dakota megafloral record. *Paleobiology* **30**(3), 347-368.
7. Sessa E.B. 2019 Polyploidy as a mechanism for surviving global change. *New Phytologist* **221**(1), 5-6.
8. Kurschner W.M., Batenburg S.J., Mander L. 2013 Aberrant Classopollis pollen reveals evidence for unreduced (2n) pollen in the conifer family Cheirolepidiaceae during the

Triassic-Jurassic transition. *Proc Biol Sci* **280**(1768), 20131708. (doi:10.1098/rspb.2013.1708).

9. Visscher H., Looy C.V., Collinson M.E., Brinkhuis H., van Konijnenburg-van Cittert J.H., Kurschner W.M., Sephton M.A. 2004 Environmental mutagenesis during the end-Permian ecological crisis. *Proc Natl Acad Sci U S A* **101**(35), 12952-12956. (doi:10.1073/pnas.0404472101).

10. Vanneste K., Sterck L., Myburg A.A., Van de Peer Y., Mizrachi E. 2015 Horsetails Are Ancient Polyploids: Evidence from Equisetum giganteum. *Plant Cell* **27**(6), 1567-1578. (doi:10.1105/tpc.15.00157).

11. Landis J.B., Soltis D.E., Li Z., Marx H.E., Barker M.S., Tank D.C., Soltis P.S. 2018 Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany* **105**(3), 348-363. (doi:10.1002/ajb2.1060).

12. Moriyama Y., Koshiba-Takeuchi K. 2018 Significance of whole-genome duplications on the emergence of evolutionary novelties. *Briefings in functional genomics*.

13. Donoghue P.C., Purnell M.A. 2005 Genome duplication, extinction and vertebrate evolution. *Trends in Ecology and Evolution* **20**(6), 312-319. (doi:10.1016/j.tree.2005.04.008).

14. Stanich N.A., Rothwell G.W., Stockey R.A. 2009 Phylogenetic diversification of Equisetum (Equisetales) as inferred from Lower Cretaceous species of British Columbia, Canada. *American Journal of Botany* **96**(7), 1289-1299. (doi:10.3732/ajb.0800381).

15. Channing A., Zamuner A., Edwards D., Guido D. 2011 Equisetum thermale sp. nov. (Equisetales) from the Jurassic San Agustín hot spring deposit, Patagonia: Anatomy, paleoecology, and inferred paleoecophysiology. *American Journal of Botany* **98**(4), 680-697. (doi:10.3732/ajb.1000211).

16. Elgorriaga A., Escapa I.H., Bomfleur B., Cúneo R., Ottone E.G. 2015 Reconstruction and Phylogenetic Significance of a New Equisetum Linnaeus Species from the Lower Jurassic of Cerro Bayo (Chubut Province, Argentina). *Ameghiniana* **52**(1), 135-152. (doi:10.5710/AMGH.15.09.2014.2758).

17. Elgorriaga A., Escapa I.H., Rothwell G.W., Tomescu A.M.F., Rubén Cúneo N. 2018 Origin of Equisetum: Evolution of horsetails (Equisetales) within the major euphyllophyte clade Sphenopsida. *American Journal of Botany* **105**(8), 1286-1303. (doi:10.1002/ajb2.1125).

18. Carruthers M., Yurchenko A.A., Augley J.J., Adams C.E., Herzyk P., Elmer K.R. 2018 De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. *BMC Genomics* **19**(1), 32. (doi:10.1186/s12864-017-4379-x).

19. Bolger A.M., Lohse M., Usadel B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114-2120.
20. Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q. 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**. (doi:10.1038/nbt.1883).
21. Fu L., Niu B., Zhu Z., Wu S., Li W. 2012 CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(23), 3150-3152.
22. Haas B., Papanicolaou A. 2012 Transdecoder. (
23. Zwaenepoel A., Van de Peer Y. 2019 wgd: simple command line tools for the analysis of ancient whole genome duplications. *Bioinformatics*.
24. Yang Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(8), 1586-1591. (doi:10.1093/molbev/msm088).
25. Edgar R.C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**. (doi:10.1093/nar/gkh340).
26. Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**(3), 307-321.
27. Emms D.M., Kelly S. 2015 OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**(1), 157. (doi:10.1186/s13059-015-0721-2).
28. Capella-Gutierrez S., Silla-Martinez J.M., Gabaldon T. 2009 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**(15), 1972-1973. (doi:10.1093/bioinformatics/btp348).
29. Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015 IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**(1), 268-274. (doi:10.1093/molbev/msu300).
30. Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018 UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**(2), 518-522. (doi:10.1093/molbev/msx281).
31. Puttick M.N. 2019 MCMCtreeR: functions to prepare MCMCtree analyses and visualise posterior ages on trees. *Bioinformatics*. (doi:10.1093/bioinformatics/btz554).
32. dos Reis M., Donoghue P.C.J., Yang Z. 2016 Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics* **17**(2), 71-80. (doi:10.1038/nrg.2015.8).

33. Morris J.L., Puttick M.N., Clark J.W., Edwards D., Kenrick P., Pressel S., Wellman C.H., Yang Z., Schneider H., Donoghue P.C. 2018 The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences* **115**(10), E2274-E2283.
34. Inoue J.G., Donoghue P.C.J., Yang Z. 2010 The impact of the representation of fossil calibrations on bayesian estimation of species divergence times. *Systematic Biology* **59**(1), 74-89.
35. dos Reis M., Yang Z. 2011 Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Molecular Biology and Evolution* **28**(7), 2161-2172. (doi:10.1093/molbev/msr045).
36. Rambaut A., Suchard M., Drummond A.J. 2014 Tracer v1. 6. (
37. Szöllősi G.J., Boussau B., Abby S.S., Tannier E., Daubin V. 2012 Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences* **109**(43), 17513-17518. (doi:10.1073/pnas.1202997109).
38. Clark J.W., Donoghue P.C. 2017 Constraining the timing of whole genome duplication in plant evolutionary history. *Proceedings of the Royal Society B: Biological Sciences* **284**(1858), 20170912.
39. Inman H.F., Bradley E.L. 1989 The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods* **18**(10), 3851-3874. (doi:10.1080/03610928908830127).
40. Ronquist F., Huelsenbeck J.P. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**. (doi:10.1093/bioinformatics/btg180).
41. Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012 A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology* **61**(6), 973-999.
42. Thorne J.L., Kishino H. 2002 Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**, 689-702.
43. Lepage T., Bryant D., Philippe H., Lartillot N. 2007 A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* **24**(12), 2669-2680.
44. Dos Reis M., Zhu T., Yang Z. 2014 The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. *Systematic biology* **63**(4), 555-565.
45. Lewis P.O. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology* **50**(6), 913-925.

544 46. Deline B., Greenwood J.M., Clark J.W., Puttick M.N., Peterson K.J., Donoghue
545 P.C.J. 2018 Evolution of metazoan morphological disparity. *Proceedings of the National*
546 *Academy of Sciences*.
547 47. Deline B. 2009 The effects of rarity and abundance distributions on measurements of
548 local morphological disparity. *Paleobiology* **35**(2), 175-189. (doi:10.1666/08028.1).
549 48. Gower J.C. 1971 A General Coefficient of Similarity and Some of Its Properties.
550 *Biometrics* **27**(4), 857-871. (doi:10.2307/2528823).
551 49. Revell L.J. 2012 phytools: An R package for phylogenetic comparative biology (and
552 other things). *Methods in Ecology and Evolution* **3**, 217-223.
553 50. Huelsenbeck J.P., Nielsen R., Bollback J.P. 2003 Stochastic mapping of
554 morphological characters. *Systematic Biology* **52**(2), 131-158.
555 51. Chartier M., Löfstrand S., von Balthazar M., Gerber S., Jabbour F., Sauquet H.,
556 Schönenberger J. 2017 How (much) do flowers vary? Unbalanced disparity among flower
557 functional modules and a mosaic pattern of morphospace occupation in the order Ericales.
558 *Proceedings of the Royal Society B: Biological Sciences* **284**(1852), 20170066.
559 (doi:10.1098/rspb.2017.0066).
560 52. Guillerme T., Cooper N., Smith A. 2018 Time for a rethink: time sub-sampling
561 methods in disparity-through-time analyses. *Palaeontology* **0**(0). (doi:10.1111/pala.12364).
562 53. Bennett M., Leitch I. 2012 Plant DNA C-values database. (Royal Botanic Gardens
563 Kew.
564 54. Franks P.J., Freckleton R.P., Beaulieu J.M., Leitch I.J., Beerling D.J. 2012
565 Megacycles of atmospheric carbon dioxide concentration correlate with fossil plant genome
566 size. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**(1588), 556.
567 55. Gould R.E. 1968 Morphology of *Equisetum laterale* Phillips, 1829, and *E. bryani*i sp.
568 nov. from the Mesozoic of South-Eastern Queensland. *Australian Journal of Botany* **16**(1),
569 153-176.
570 56. Pagel M. 1999 Inferring the historical patterns of biological evolution. *Nature* **401**,
571 877. (doi:10.1038/44766).
572 57. Vanneste K., Van de Peer Y., Maere S. 2013 Inference of genome duplications from
573 age distributions revisited. *Molecular Biology and Evolution* **30**(1), 177-190.
574 (doi:10.1093/molbev/mss214).
575 58. Gould S.J., Lewontin R.C. 1979 The spandrels of San Marco and the Panglossian
576 paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B* **205**(1161), 581-598.

59. Freeling M. 2017 The distribution of ancient polyploidies in the plant phylogenetic tree is a spandrel of occasional sex. *The Plant Cell*.
60. Levin D.A., Soltis D.E. 2018 Factors promoting polyploid persistence and diversification and limiting diploid speciation during the K–Pg interlude. *Current opinion in plant biology* **42**, 1-7.
61. Nakatani Y., McLysaght A. 2019 Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proceedings of the National Academy of Sciences* **116**(6), 1816-1818.
62. Doyle J.J., Egan A.N. 2010 Dating the origins of polyploidy events. *New Phytologist* **186**(1), 73-85. (doi:10.1111/j.1469-8137.2009.03118.x).
63. Schwager E.E., Sharma P.P., Clarke T., Leite D.J., Wierschin T., Pechmann M., Akiyama-Oda Y., Esposito L., Bechsgaard J., Bilde T., et al. 2017 The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology* **15**(1), 62. (doi:10.1186/s12915-017-0399-x).
64. O'Reilly J.E., Puttick M.N., Pisani D., Donoghue P.C.J. 2017 Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology* **61**(1), 105-118. (doi:10.1111/pala.12330).
65. Puttick M.N., Reilly J.E., Tanner A.R., Fleming J.F., Clark J., Holloway L., Lozano-Fernandez J., Parry L.A., Tarver J.E., Pisani D., et al. 2017 Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B: Biological Sciences* **284**(1846).
66. Edger P.P., Heidel-Fischer H.M., Bekaert M., Rota J., Glöckner G., Platts A.E., Heckel D.G., Der J.P., Wafula E.K., Tang M., et al. 2015 The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences* **112**(27), 8362-8366. (doi:10.1073/pnas.1503926112).
67. Taylor E.L., Taylor T.N., Krings M. 2009 *Paleobotany: the biology and evolution of fossil plants*, Academic Press.
68. Stewart W.N., Stewart W.N., Rothwell G.W. 1993 *Paleobotany and the evolution of plants*, Cambridge University Press.
69. Puttick M.N., Clark J., Donoghue P.C.J. 2015 Size is not everything: rates of genome size evolution, not C-value, correlate with speciation in angiosperms. *Proceedings of the Royal Society B: Biological Sciences* **282**(1820).

70. Leitch I.J., Bennett M.D. 2004 Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* **82**(4), 651-663. (doi:DOI 10.1111/j.1095-8312.2004.00349.x).

71. Clark J., Hidalgo O., Pellicer J., Liu H., Marquardt J., Robert Y., Christenhusz M., Zhang S., Gibby M., Leitch I.J., et al. 2016 Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytologist* **210**(3), 1072-1082. (doi:10.1111/nph.13833).

Figure 1. Node-averaged rates of synonymous substitution (*Ks*) between paralogous pairs for A) *Equisetum diffusum* and B) *Equisetum hyemale*. Components among the distributions were fitted using the function *gmm()* in the *wgd* pipeline.

Figure 2. A histogram showing the combined posterior distribution of ages for the duplication node among 240 gene families containing the signal of a gene duplication event in *Equisetum.* Two clusters are defined using mixture models.

Figure 3. Inferred age of the whole genome duplication (WGD) event in *Equisetum*. Multi-copy gene families were concatenated to inform a molecular clock analysis for each putative

WGD event. The 95% HPD is shown for each speciation node in blue, with the duplication events in red.

Figure 4. Total evidence phylogeny of extinct and extant Equisetales. The tree was constructed using Bayesian analysis of phenotypic and molecular data with the ages of the fossils as tip calibrations and nodes calibrated using estimates from the molecular analysis. Rates of phenotypic evolution (low rates in blue, high rates in red) are from the mean effective branch rates from a posterior sample of 1000 trees estimated morphological data alone. High rates are shown in text next to branches. The position of each putative WGD is shown on the tree.

Figure 5. Phenotypic evolution within the Equisetales. A) An empirical phylomorphospace showing the distribution of disparity within the order. The distances between taxa were calculated using Gower's index and ordinated using non-metric multidimensional scaling (NMDS). Character states for all ancestral nodes were reconstructed and were projected into the morphospace with the tree. Convex hulls were fitted around each lineage. Colours correspond to different lineages. B) The comparative morphospace occupation of extant and fossil Equisetales. C) The evolution of disparity (Sum Of Variances) through time estimated from the distance matrix.

Figure 6. The reconstruction of ancestral genome size across the Equisetales. The genome size was reconstructed based on both extant and fossil 1C-value estimates. The reconstructed size is shown at each node, with the width of the circle proportional to the 1C-value. The middle circle represents the mean estimate, while the small and large circles represent the lower and upper 95% HPD values, respectively. Branches are coloured to show the evolution of large (red) and small (blue) genome sizes.